VIRUS EVOLUTION

Virus Evolution, 2023, **9(2)**, 1–8
DOI: https://doi.org/10.1093/ve/vead040
Advance Access Publication 29 June 2023
Research Article

Evolutionary changes in the number of dissociable amino acids on spike proteins and nucleoproteins of SARS-CoV-2 variants

Anže Božič[1,†] and Rudolf Podgornik[2,3,4,5,6,*]

[1]Department of Theoretical Physics, Jožef Stefan Institute, Jamova 39, Ljubljana SI-1000, Slovenia, [2]School of Physical Sciences, University of Chinese Academy of Sciences, No. 19A Yuquan Road, Shijingshan District, Beijing 100049, China, [3]Kavli Institute for Theoretical Sciences, University of Chinese Academy of Sciences, No. 3 Nanyitiao, Zhongguancun, Haidian District, Beijing 100049, China, [4]CAS Key Laboratory of Soft Matter Physics, Institute of Physics, Chinese Academy of Sciences, No. 8 3rd South Street, Zhongguancun, Haidian District, Beijing 100190, China, [5]Wenzhou Institute of the University of Chinese Academy of Sciences, No.1 Jinlian Road, Wenzhou, Zhejiang 325001, China and [6]Department of Physics, Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, Ljubljana SI-1000, Slovenia
[†]https://orcid.org/0000-0001-6304-6637
*Corresponding author: E-mail: podgornikrudolf@ucas.ac.cn

Abstract

The spike protein of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is responsible for target recognition, cellular entry, and endosomal escape of the virus. At the same time, it is the part of the virus that exhibits the greatest sequence variation across the many variants which have emerged during its evolution. Recent studies have indicated that with progressive lineage emergence, the positive charge on the spike protein has been increasing, with certain positively charged amino acids (AAs) improving the binding of the spike protein to cell receptors. We have performed a detailed analysis of dissociable AAs of more than 1400 different SARS-CoV-2 lineages, which confirms these observations while suggesting that this progression has reached a plateau with Omicron and its subvariants and that the positive charge is not increasing further. Analysis of the nucleocapsid protein shows no similar increase in positive charge with novel variants, which further indicates that positive charge of the spike protein is being evolutionarily selected for. Furthermore, comparison with the spike proteins of known coronaviruses shows that already the wild-type SARS-CoV-2 spike protein carries an unusually large amount of positively charged AAs when compared to most other betacoronaviruses. Our study sheds light on the evolutionary changes in the number of dissociable AAs on the spike protein of SARS-CoV-2, complementing existing studies and providing a stepping stone towards a better understanding of the relationship between the spike protein charge and viral infectivity and transmissibility.

Keywords: severe acute respiratory syndrome coronavirus 2; spike protein; nucleoprotein; dissociable amino acids; protein charge evolution.

Introduction

The coronavirus disease 2019 (COVID-19) pandemic caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) gave rise to an unprecedented global effort to collect and share the information of the virus' ongoing evolution (Moorthy et al. 2020; Pratt, Bull, and Med 2021) with its emerging stream of new variants with varying degrees of disease severity, transmissibility, and other traits (Singh and Yi 2021; Telenti, Hodcroft, and Robertson 2022). Consequently, this large amount of data enables the study of how the virus has been changing since it was first detected in humans and the discovery of those mutations which made it adapt further, with a particular focus on the designated variants of interest (VOIs) and variants of concern (VOCs) (Magazine et al. 2022; Obermeyer et al. 2022; Bloom and Neher 2023; Carabelli et al. 2023). The spike (S) protein of the virus—responsible for target recognition, cellular entry, and endosomal escape (Huang et al. 2020)—in particular exhibits the greatest sequence variation, not only across different SARS-CoV-2 variants but also in coronaviruses in general (Cavanagh 2005; Harvey et al. 2021). As an example, the highly transmissible Omicron variant has fifteen mutations solely in its receptor-binding domain (RBD), the part of the S protein that binds to angiotensin-converting enzyme 2 (ACE2) receptor. The high variability of the S protein across different variants has also led to recent efforts in studying whether VOCs could be identified from the sequence of the S protein alone, which would make lineage assignment relatively easier compared to the use of complete genome sequences (O'Toole et al. 2022).

Numerous experimental, computational, and bioinformatics studies have analysed the influence of the observed mutations in the S protein on the viral function. Some of the mutations have been linked to an increased transmissibility of the virus, while

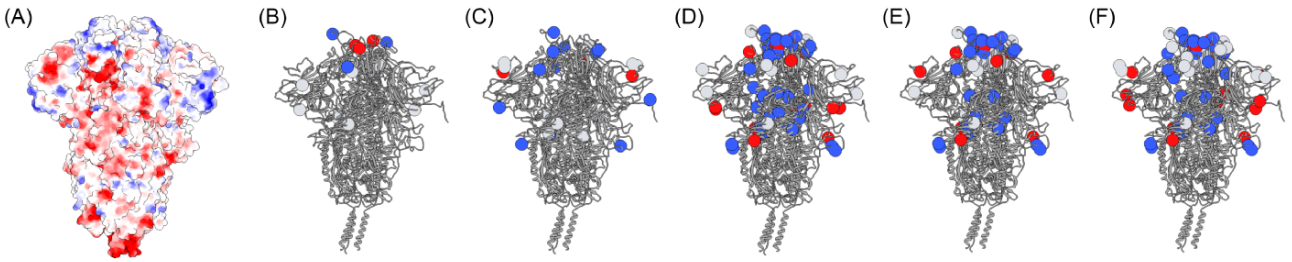

Figure 1. (A) Coulombic surface colouring of the WT SARS-CoV-2 S protein (Protein Data Bank (PDB) entry: 7FB0), showing the positions of positively and negatively charged AAs (blue and red, respectively) on the S protein surface. The actual charge distribution on the S protein is a more complex question, which depends, among other things, on the bathing solvent conditions, such as electrolyte concentration and pH, and local values of AA dissociation constants (Javidpour et al. 2021; Kim et al. 2022). Mutations of dissociable AAs on the S protein of selected SARS-CoV-2 VOCs: beta (B), delta (C), Omicron BA.1 (D), Omicron BA.4 (E), and Omicron XBB (F). Only those mutations that lead to either a gain or a loss of a dissociable AA are shown (Hodcroft 2021): gain of a positively charged AA (blue), gain of a negatively charged AA (red), and loss of either positively or negatively charged AA (light gray). Mutations which replace one dissociable AA with another of the same charge type are not represented. Note that these images show only the positions of the mutations of dissociable AAs and do not depict the changes in charge on the S protein. The complete PDB structure of the S protein in (B)–(F) was obtained from Kim et al. (2022). All the images were drawn using UCSF Chimera (Pettersen et al. 2004).

others have been linked to changes in its binding to ACE2 (Jawad et al. 2022; Obermeyer et al. 2022; Sang et al. 2022; Zhang et al. 2022). A particularly interesting observation, made as the COVID-19 pandemic progressed, concerns the tendency of the mutations in the S protein of different VOCs to increase the number of positively charged amino acid (AA) residues in it (Pawłowski 2021) (illustrated on a few examples in Fig. 1). This observation is further supported by a more recent analysis of the charge on the S protein of major SARS-CoV-2 lineages, which identified a striking change in the S protein charge with the evolution of the virus (Cotten and Phan 2023). Molecular dynamics and large-scale ab initio studies also found that the increase in the (partial) charge of S protein RBD of some VOCs increases the binding to ACE2 and other cell surface receptors (Adhikari et al. 2022a, 2022b, Nie et al. 2022; Gan et al. 2022; Kim et al. 2022) and that the total charge of the RBD might be a simple predictor for the RBD–ACE2 binding affinity based on the data obtained for main VOCs (Barroso da Silva, Giron, and Laaksonen 2022). That charge might influence viral infectivity has also been shown in avian influenza viruses, where hemagglutinin proteins from low and high pathogenic strains exhibit clear difference in their surface charge (Baggio, Filippini, and Righetto 2023). These observations go hand in hand with a large body of studies showing the electrostatic interactions to be of enormous importance both in biological systems, in general (Holm, Kékicheff, and Podgornik 2000; Zhou and Pang 2018), as well as in viruses, in particular (Šiber, Božič, and Podgornik 2012; Zandi et al. 2020).

To study the effect of S protein charge on, for instance, its binding affinity requires significant computational effort and more often than not involves a certain degree of approximation (Javidpour et al. 2021; Barroso da Silva, Giron, and Laaksonen 2022; Kim et al. 2022), unless one resorts to quantum-level calculations (Ching et al. 2023). On the other hand, the amount of dissociable AAs provides a good estimate for the total charge of the S protein and importantly enables a broad study of different SARS-CoV-2 lineages (Pawłowski 2021; Cotten and Phan 2023). In this work, we use the large amount of available data on different SARS-CoV-2 lineages that have emerged over the past 3 years of the pandemic to examine in detail how the number of dissociable AAs on the S protein—as a proxy for its total charge—has changed with increasing lineage divergence and evolution of the virus. We observe several different clusters corresponding to the emergence of different variants. We indeed observe a general tendency toward an increase in the total number of positively charged AAs on the S protein, a tendency that, however, seems to have recently plateaued. For comparison, we perform the same analysis on the SARS-CoV-2 nucleoprotein (N protein) to show that the evolutionary preference for positive charge is specific to the S protein. We additionally examine the available data on S and N proteins in known coronaviruses to frame our results in a wider context. In this way, our work complements the existing studies on the importance of S protein charge for its interaction with the environment, at the same time adding several novel observations regarding the preference for particular dissociable AAs in different variants and the saturation of their number with Omicron VOC and its many subvariants.

Results

Dissociable AAs on S and N proteins show different amounts of change with lineage divergence

In general, there are six AAs that can (de)protonate and thus acquire charge: three of them can be positively charged (arginine (ARG), lysine (LYS), and histidine (HIS)), whereas three of them can be negatively charged (aspartic acid (ASP), glutamic acid (GLU), and tyrosine (TYR))—see also the Methods section. Figure 2 shows how the average number of different dissociable AAs on the S and N proteins of 1421 SARS-CoV-2 variants changes with increasing average lineage divergence from the wild-type (WT) genome. Since the length of the S protein is approximately three times the length of the N protein (≈1,270 AA compared to ≈415 AA, respectively), it is not surprising that the number of dissociable AAs on the S protein is, in general, much larger than the number of dissociable AAs on the N protein. One can, nonetheless, observe very different trends in their numbers as lineage divergence increases. For instance, the number of positively charged LYS AAs on the S protein tends to steadily increase with lineage divergence, with a peak number of around sixty-seven with the Omicron subvariants BA.1, BA.3, and the recombinant XD (Fig. 2A). However, this number slightly decreases afterward for the more divergent subvariants BA.4, BA.5, and the recombinant XBB—see also Fig. 3 and Table 1. The number of HIS also increases with increasing divergence, albeit it does so only at relatively highly divergent lineages. The changes in the number of negatively charged AAs are less prominent, with perhaps the exception of TYR, whose number is slightly increased with the more divergent Omicron subvariants.

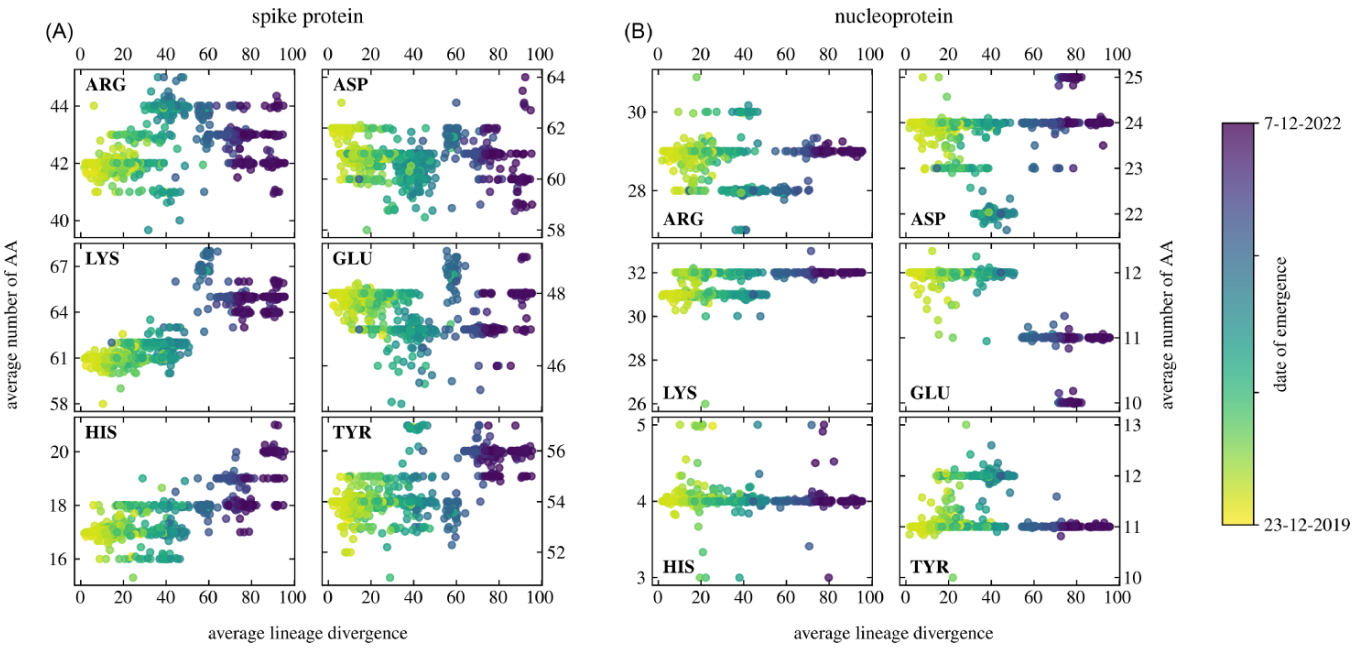

Figure 2. Average change in the number of dissociable AAs on (A) the S protein and (B) the N protein of 1,421 different SARS-CoV-2 lineages as a function of average lineage divergence. The left column of each panel shows positively charged AAs (ARG, LYS, and HIS), and the right column of each panel shows negatively charged AAs (ASP, GLU, and TYR). Datapoint colours correspond to the earliest known isolation date of the lineage as indicated by the colour bar.

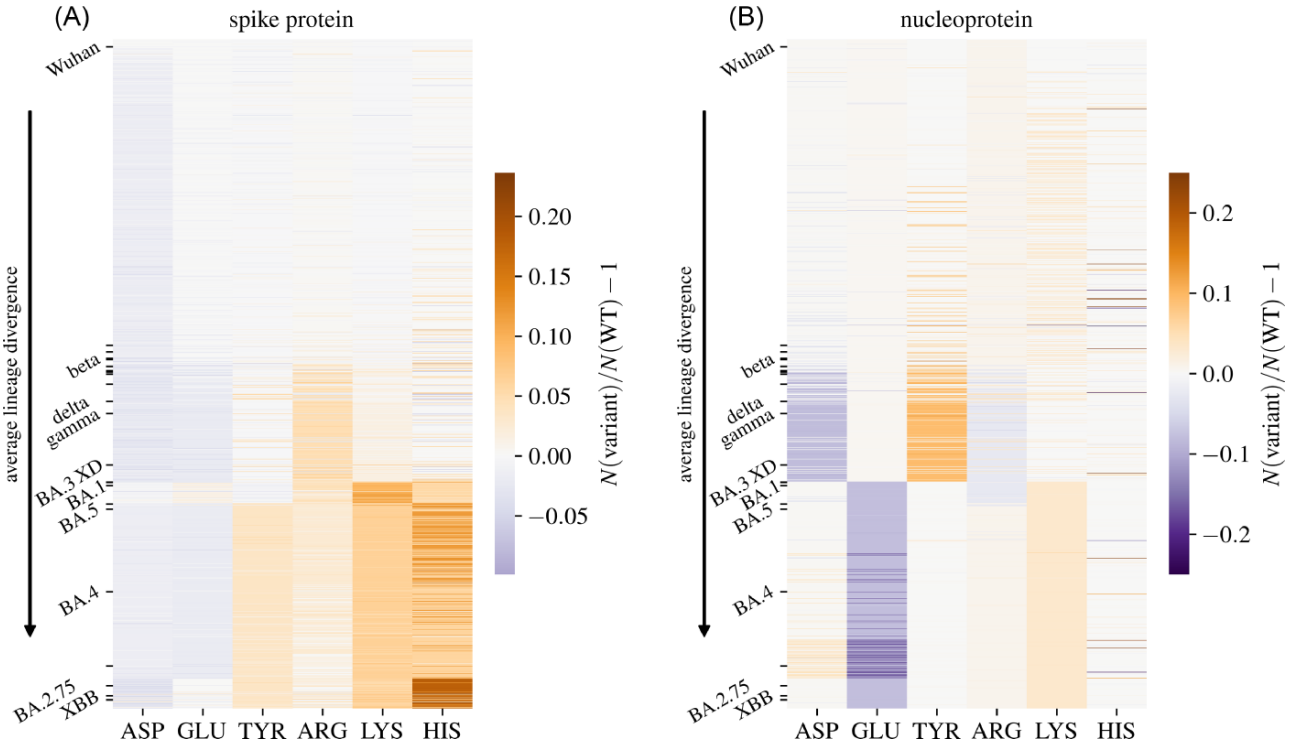

Figure 3. Heatmap of the relative change of the average number of dissociable AAs on (A) the S protein and (B) the N protein of 1,421 SARS-CoV-2 lineages compared to the WT. Lineages are sorted in the order of increasing divergence. Ticks and labels on the y axes mark select VOIs and VOCs along the divergence progression (cf. Table 1).

In contrast to the S protein, the number of dissociable AAs on the N protein does not show any significant increases or decreases with lineage divergence (Fig. 2B). Here, the more interesting observation is that the number of certain AAs, such as TYR and LYS, occupied two distinct values in the early variants. As the lineages began to diverge more, only one of the values is selected for (a lower number of TYR and a higher number of LYS). Nonetheless, the number of dissociable AAs on the N protein shows far fewer changes with divergent SARS-CoV-2 lineages compared to their number in the S protein.

Table 1. Average lineage divergence of selected SARS-CoV-2 VOCs and VOIs.

Variant	Pango lineage	Average divergence
Wuhan	B	5.02
Iota	B.1.526	25.2
Zeta	P.2	25.8
Beta	B.1.351	27.7
Epsilon	B.1.429	28.1
Eta	B.1.525	30.5
Kappa	B.1.617.1	32.5
Theta	P.3	33.0
Lambda	C.37	33.8
Mu	B.1.621	36.9
Delta	B.1.617.2	38.8
Gamma	P.1	39.7
Omicron XD	XD	44.4
Omicron BA.3	BA.3	54.9
Omicron BA.1	BA.1	56.1
Omicron XE	XE	63.7
Omicron BA.5	BA.5	68.3
Omicron BA.4	BA.4	73.5
Omicron XAY	XAY	79.0
Omicron BA.2.75	BA.2.75	88.7
Omicron XBB	XBB	91.1
Omicron BJ.1	BJ.1	92.0

Different positively charged AA types are preferred with lineage divergence

As Fig. 2 shows, different dissociable AAs show different patterns of change with the evolution of SARS-CoV-2 and the emergence of new lineages. In order to see whether any patterns can be observed between different lineages with respect to their preference for a particular AA type, Fig. 3 shows a heatmap of the relative changes in the number of dissociable AAs on the S and N proteins from different SARS-CoV-2 lineages compared to the WT, with select VOCs and VOIs marked along the divergence progression. Looking at the S protein first (Fig. 3A), we can observe that the first positively charged AA residue to show an increase is ARG, which reaches a slight peak in a cluster of variants that covers Delta and Gamma variants (cf. also Table 1). The number of ARG, however, decreases again for more divergent variants, and in its place, the number of LYS is significantly increased, in a cluster covering Omicron subvariants BA.1, BA.3, and XD. While the number of LYS also remains high for later, more divergent subvariants, its number, nonetheless, drops in comparison to this cluster. Interestingly, the most divergent variants, including the Omicron subvariants B.2.75, BJ.1, and the recombinant XBB, show an increased number of HIS residues, which are only fractionally charged at physiological pH. The numbers of negatively charged AAs on the S protein change less drastically, but have a tendency toward a slight decrease in the more divergent lineages, with the notable exception of TYR, the number of which is increased in most Omicron subvariants.

While the overall number of dissociable AAs on the N protein changes less drastically (Fig. 3B), some trends can nonetheless be observed. Variants such as Delta and Gamma show an increase in the number of TYR and a decrease in the number of ASP. On the other hand, these changes are absent in the Omicron subvariants, where the most notable observation is the decrease in the number of GLU residues. We again observe that as lineage divergence increases, the number of positively charged AAs on the N protein shows a much lesser degree of change compared to their counterparts on the S protein. Figures 3 and 2 together make it clear that the evolutionary preference for the increase in the amount of positively charged AAs is particular to the S protein of SARS-CoV-2, and no similar effect can be observed for the number of dissociable AAs on the N protein.

Number of positively charged AAs on the S protein has plateaued with Omicron subvariants

As already mentioned, the total number of dissociable AAs can serve as a proxy for the amount of charge on the S and N proteins. Figure 4 separates the contributions of the positively (ARG, LYS, and HIS) and negatively (ASP, GLU, and TYR) charged AAs and shows how their total number changes with the increasing divergence of SARS-CoV-2 lineages from the WT genome. Similar to what has been observed previously (Pawłowski 2021; Cotten and Phan 2023), the number of positively charged AAs on the S protein increases with lineage divergence, while the number of negatively charged AAs remains rather steady or even decreases slightly for highly divergent lineages (Fig. 4A). This indeed indirectly implies that the overall charge on the S protein has been becoming more positive with increasing lineage divergence. However, we can also observe that the total number of positively charged AAs appears to have reached a plateau with the Omicron variant, with only small changes in the amount of positively charged AAs observed between its different subvariants.

The changes in the total number of positively and negatively charged AAs on the N protein as lineage divergence increases are, on the other hand, again far smaller compared to those of the S protein (Fig. 4B). However, we can also observe that the most divergent (and recently emerged) lineages seem to clearly prefer a version of the N protein with slightly fewer negatively charged AAs and slightly more positively charged AAs compared to the WT, albeit with a significantly smaller variation in their number compared to the less divergent lineages. This, combined with the observations of the number of charged AAs on the S protein (Fig. 4A), implies that the number of dissociable AAs on both the S and N proteins has reached an 'equilibrium' where any further significant changes appear to be less likely.

SARS-CoV-2 has more positively charged AAs than other known (beta)coronaviruses

Compared to the abundance of data on SARS-CoV-2, there is much less information available regarding the evolution of the S proteins of other coronaviruses. Nonetheless, we can compare the number of dissociable AAs on the S protein of most currently known coronaviruses based on their reference sequences. Figure 5 thus shows the comparison of the number of positively and negatively charged AAs on the S proteins of the reference strains of forty-six different coronaviruses (see the Methods section), together with twenty-two VOCs and VOIs of SARS-CoV-2 (Table 1). In general, the S proteins of coronaviruses tend to have a larger number of GLU and ASP AAs compared to the number of ARG and LYS AAs (Fig. 5A), as well as a large amount of TYR (Fig. 5B), which indicates that the overall charge of the S protein is negative. There are a few exceptions that have approximately the same number of GLU and ASP AAs compared to ARG and LYS, including human coronavirus NL63, suncus murinus coronavirus X74, two avian coronaviruses (IBV and 9203), and two bat coronaviruses (Rhinolophus bat coronavirus HKU2 and Rousettus bat coronavirus HKU9). Compared to most other betacoronaviruses—and, in fact, most

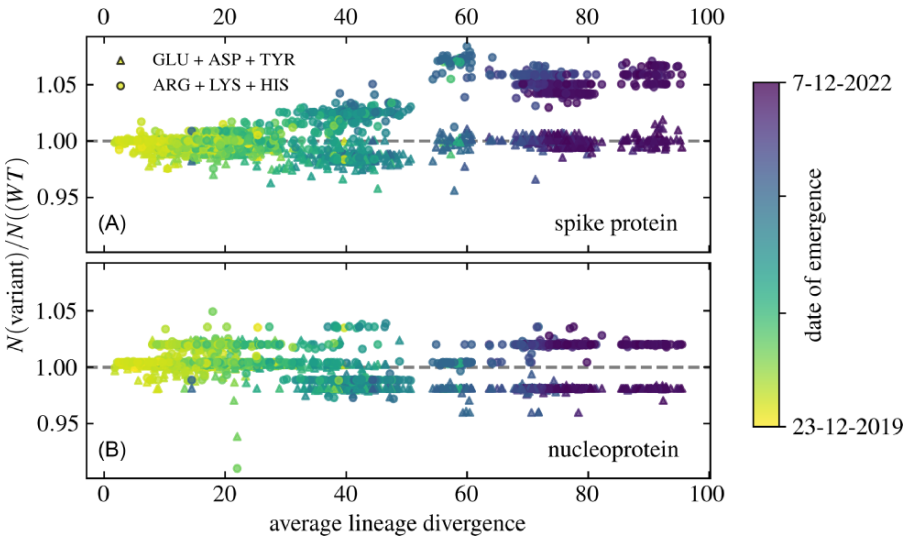

Figure 4. Change in the average total number of AAs on (A) the S protein and (B) the N protein of 1,421 different SARS-CoV-2 lineages compared to the WT, shown as a function of the average lineage divergence. AAs are grouped by their ionizability: ASP, GLU, and TYR (negative), on the one hand, and ARG, LYS, and HIS (positive), on the other.

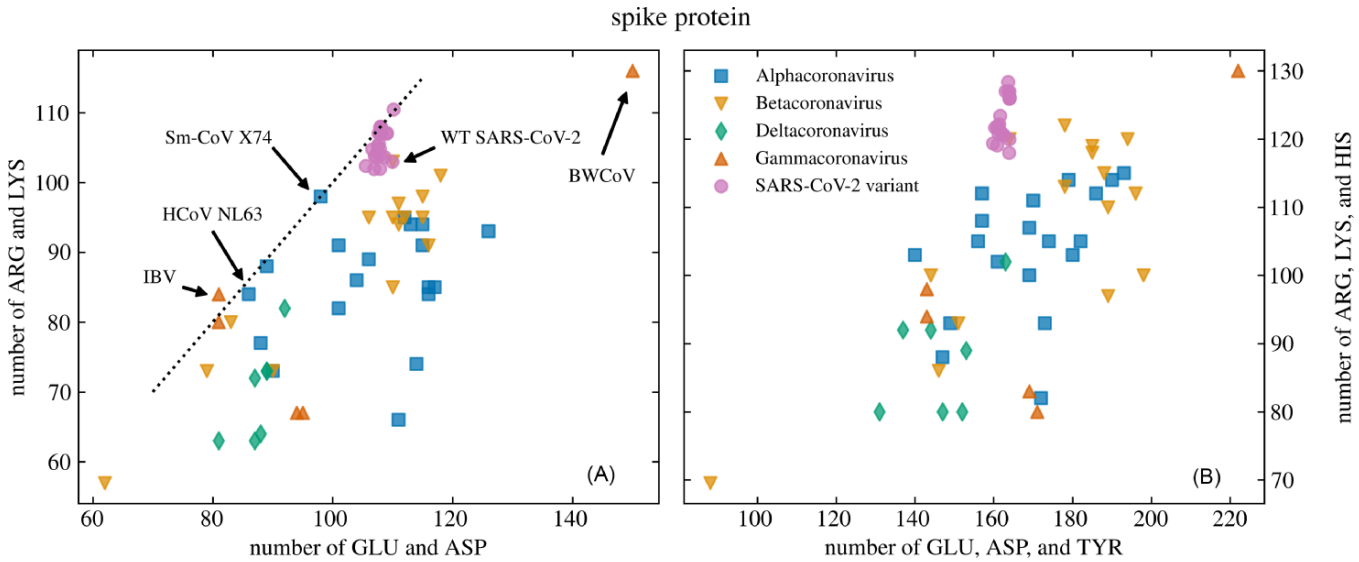

Figure 5. Comparison of the number of negatively and positively charged AAs on the S proteins of different coronaviruses. The comparison is shown for forty-six different viruses from the Coronaviridae family and for twenty-two variants of SARS-CoV-2 (Table 1). (A) The number of ARG and LYS residues compared to the number of ASP and GLU residues; (B) the number of ARG, LYS, and HIS residues compared to the number of ASP, GLU, and TYR residues. The dotted line in (A) shows where the number of negatively and positively charged AAs is equal. Arrows mark some of the coronaviruses mentioned in the main text.

other coronaviruses in general—both WT SARS-CoV-2 and its variants have a larger number of ARG and LYS AAs on their S proteins. One notable exception is SARS-CoV, which has a similar amount of positively and negatively charged AAs as WT SARS-CoV-2. In general, Fig. 5 shows that the S protein of WT SARS-CoV-2 already has fewer negatively charged AAs and more positively charged AAs than other betacoronaviruses and that with lineage divergence, it is the number of positively charged AAs that has increased even further.

Discussion and Conclusions

By analysing the number of dissociable AAs on the S and N proteins of more than 1,400 different SARS-CoV-2 lineages, we have shown that there is an overall tendency towards an increase in the number of positive AAs on the S protein with increasing lineage divergence, an effect that is not readily observed for the N protein (Fig. 3). While the number of dissociable AAs on a protein can only be considered a proxy for its total charge, which is in fact a result of many different local and global factors (Adhikari et al. 2022a), our observations confirm that a more positively charged S protein of SARS-CoV-2 is evolutionarily favourable. This result is in line with previous studies (Pawłowski 2021; Barroso da Silva, Giron, and Laaksonen 2022; Cotten and Phan 2023) which have shown that the charge on the S protein has been increasing ever since the start of the pandemic. However, we have also shown that the overall number of positively charged AAs has seemingly reached a plateau with Omicron and its subvariants (Figs 4 and 1).

Further observation of changes in the number of dissociable AAs in emerging variants as SARS-CoV-2 continues to evolve will show whether this plateau is only temporary or whether the positive charge on the S protein has reached its peak. On the other hand, our observation that the positive charge on the N protein, which plays a role in condensing the RNA genome (Dinesh et al. 2020; Cascarina and Ross 2022; Li and Zandi 2022), has changed only little with the evolution of SARS-CoV-2 might indicate that it is already optimized. We also note that while a similar analysis can be carried out on other structural proteins of SARS-CoV-2, such as the envelope and membrane proteins, these are significantly smaller than the S and N proteins and have far fewer dissociable AAs to start with. Consequently, any evolutionary changes in the number of dissociable AAs occur on a much smaller scale—on the order of a change in a single AA residue—and are thus difficult to compare.

Detailed inspection of the exact AA composition of the S protein in different SARS-CoV-2 lineages showed that the main contribution to positive charge comes from additional LYS residues, which reached the largest number with BA.1 and BA.3 Omicron subvariants and the XD recombinant variant (Fig. 2A). Interestingly, the most divergent lineages, including the Omicron subvariants BA.2.75 and BJ.1 as well as the XBB recombinant variant, also show a significant increase in the number of HIS residues, which is only fractionally charged at neutral pH. Comparison with the changes in the number of dissociable AAs on the N proteins of different SARS-CoV-2 lineages (Fig. 2B) shows that while there is a slight tendency for an increase in charge with increasing lineage divergence, this occurs to a far lesser extent. We argue that this is an additional confirmation that charge plays an important role in the function(s) of the S protein and that the observed increase in the number of positively charged AAs on it is not a general effect that would occur in any viral protein as lineages continue to diverge.

Numerous studies have demonstrated how individual AA mutations that increase the local charge in the RBD of the S protein change its interaction with the ACE2 receptor (which is not the only receptor that binds to the S protein of SARS-CoV-2 (Loo et al. 2023)). Even here, the question of positive charge is a complex issue: on the one hand, individual mutations that increase positive charge can reduce the binding affinity, as they are incompatible with particular LYS residues on the receptor (Zhang et al. 2022). On the other hand, the absence of charge-increasing Q493R mutation in the BA.4 and BA.5 Omicron subvariants results in a significantly weaker binding affinity to ACE2 compared to Omicron subvariants BA.1, BA.2, and BA.3 (Sang et al. 2022). Charge can also be an important factor in other interactions, as it can disrupt the hydrogen bond network and thus influence the overall interaction (Ching et al. 2023), and it has the potential to diminish antibody binding (Harvey et al. 2021). Our results, together with previous studies (Pawłowski 2021; Cotten and Phan 2023), clearly demonstrate that the evolutionary progress of SARS-CoV-2 lineages favours an *overall* increase in the positive charge of the S protein, making it stand out in this respect from among other known betacoronaviruses (Fig. 5). These changes likely have an effect that is greater than the contributions of individual AA mutations themselves: for instance, anionic, negatively charged lipids represent a dominant fraction of charged lipid species in biological membranes (Galassi and Wilke 2021), and consequently, their role in the interaction between proteins and membranes is of great biological interest (Pöyry and Vattulainen 2016). More positively charged proteins would then interact more strongly with the membranes, consequently making the positive charge mutations more desirable. We hope that further studies can elucidate how the observed increase in the overall positive charge of the S protein benefits viral infectivity, transmissibility, and other traits.

Methods

Data collection

SARS-CoV-2 variants

We obtained a list of SARS-CoV-2 Pango lineages from the CoV-Lineages.org lineage report (O'Toole et al. 2021) on 24 November 2022. These lineages were used as an input to download virus genomic and protein data from NCBI Virus (Hatcher et al. 2017) using the provided command line tools; the data were downloaded between 30 November 2022 and 5 January 2023. We used the accompanying annotations to obtain the isolate collection dates and kept the earliest record with an available full date of collection (i.e. year, month, and day) as the timepoint of the lineage 'emergence' for use in our analysis.

For each Pango lineage, we furthermore obtained the information on lineage divergence—the number of nucleotide changes (mutations) in *the entire genome* relative to the root of the phylogenetic tree, i.e. the start of the outbreak—from the global SARS-CoV-2 data available on Nextstrain (Hadfield et al. 2018). We have selected only those entries with a genome coverage of >99 per cent and extracted their lineage divergence and the number of mutations. Since individual entries within a Pango lineage can still exhibit small differences in their lineage divergence from the WT genome, we averaged over them to obtain the average lineage divergence for each Pango lineage. To allow for an easier interpretation of our results, we list in Table 1 a comparison of the average lineage divergence of selected VOCs and VOIs used in our analysis.

As the last selection step, we retained only those Pango lineages whose downloaded protein fasta file was not empty. We used these protein data to obtain the number of dissociable AAs on the S and N proteins. The numbers of dissociable AAs were then averaged over all available protein data for a given Pango lineage. The final number of analysed SARS-CoV-2 Pango lineages for which the entirety of the data described earlier was attainable is $N = 1421$.

Coronaviridae

As a point of comparison, we also examined the number of dissociable AAs on the S and N proteins of known coronaviruses. We used the coronaviruses listed in the most recent Virus Metadata Resource (2 December 2022) issued by the International Committee on Taxonomy of Viruses (International Committee on Taxonomy of Viruses 2021) and limited ourselves to the genomes of those viruses with available REFSEQ accession numbers. We used these to download the representative genome and protein fasta files from NCBI Virus (Hatcher et al. 2017). Due to the large amount of variation in the annotations across the different coronavirus datasets, we limited ourselves solely to the REFSEQ genomes and neglected any information on different strains and variants. These data were downloaded on 28 January 2023. We then followed the same procedure as with SARS-CoV-2 variants to obtain the number of dissociable AAs on the S and N proteins of different coronaviruses.

Our final dataset of coronaviruses includes forty-six different species; of these, one is a repetition of WT SARS-CoV-2. The dataset also includes SARS-CoV, Middle East respiratory syndrome-related coronavirus, and all other four known human coronaviruses. In general, the dataset comprises

nineteen species from the genus Alphacoronavirus, fifteen species from the genus Betacoronavirus, five species from the genus Gammacoronavirus, and seven species from the genus Deltacoronavirus.

Dissociable AAs

We analysed the S and N proteins of SARS-CoV-2 variants and coronaviruses to obtain the (average) numbers of six dissociable AAs: GLU, ASP, TYR, ARG, LYS, and HIS. We used Biopython to parse the protein fasta files and count the number of dissociable AAs on the S and N proteins. Three of the six AAs can carry positive charge (ARG, LYS, and HIS), while the other three can carry negative charge (ASP, GLU, and TYR) (Lide 2013). HIS typically carries a relatively small fractional charge at physiological pH, while TYR starts to acquire charge only at very basic pH; the importance of the latter in charge-mediated interactions has been demonstrated in recent studies (Barroso da Silva, Giron, and Laaksonen 2022). In our analysis, we did not consider cysteine, which has a thiol with a functional end group that is a very weak acid and is usually not considered to be an acid at all (Nap et al. 2014; Božič and Podgornik 2017).

Data availability

All the data presented in this work and a detailed description of data collection are openly available in OSF at https://osf.io/78b3f/, reference number 78B3F.

Acknowledgements

A.B. acknowledges support of Slovenian Research Agency (ARRS) under contract no. P1-0055. R.P. acknowledges funding from the Key Project under contract no. 12034019 of the National Natural Science Foundation of China.

Conflict of interest: None declared.

References

Adhikari, P. et al. (2022a) 'Mutations of Omicron Variant at the Interface of the Receptor Domain Motif and Human Angiotensin-Converting Enzyme-2', *International Journal of Molecular Sciences*, 23: 2870.

Baggio, G., Filippini, F., and Righetto, I. (2023) 'Comparative Surface Electrostatics and Normal Mode Analysis of High and Low Pathogenic H7N7 Avian Influenza Viruses', *Viruses*, 15: 305.

Barroso da Silva, F. L., Giron, C. C., and Laaksonen, A. (2022) 'Electrostatic Features for the Receptor Binding Domain of SARS-COV-2 Wildtype and its Variants. Compass to the Severity of the Future Variants with the Charge-Rule', *The Journal of Physical Chemistry. B*, 126: 6835–52.

Bloom, J. D., and Neher, R. A. (2023) 'Fitness Effects of Mutations to SARS-CoV-2 Proteins', *bioRxiv* 2023.01.30.526314.

Božič, A., and Podgornik, R. (2017) 'pH Dependence of Charge Multipole Moments in Proteins', *Biophysical Journal*, 113: 1454–65.

Carabelli, A. M. et al. (2023) 'SARS-CoV-2 Variant Biology: Immune Escape, Transmission and Fitness', *Nature Reviews Microbiology*, 21: 162–77.

Cascarina, S. M., and Ross, E. D. (2022) 'Phase Separation by the SARS-CoV-2 Nucleocapsid Protein: Consensus and Open Questions', *Journal of Biological Chemistry*, 298: 101677.

Cavanagh, D. (2005) 'Coronaviruses with Special Emphasis on First Insights Concerning SARS', In: Schmidt, A., Wolff, M., and Weber, O. (eds). pp. 1–54. Basel: Birkhäuser Verlag.

Ching, W.-Y. et al. (2023) 'Towards Quantum-Chemical Level Calculations of SARS-CoV-2 Spike Protein Variants of Concern by First Principles Density Functional Theory', *Biomedicines*, 11: 517.

Cotten, M., and Phan, M. V. (2023) 'Evolution of Increased Positive Charge on the SARS-CoV-2 Spike Protein may be Adaptation to Human Transmission', *iScience*, 26: 106230.

Dinesh, D. C. et al. (2020) 'Structural Basis of RNA Recognition by the SARS-CoV-2 Nucleocapsid Phosphoprotein', *PLOS Pathogens*, 16: e1009100.

Galassi, V. V., and Wilke, N. (2021) 'On the Coupling Between Mechanical Properties and Electrostatics in Biological Membranes', *Membranes*, 11: 478.

Gan, H. H. et al. (2022) 'Omicron Spike Protein has a Positive Electrostatic Surface that Promotes ACE2 Recognition and Antibody Escape', *Frontiers in Virology*, 2: 894531.

Hadfield, J. et al. (2018) 'Nextstrain: Real-Time Tracking of Pathogen Evolution', *Bioinformatics*, 34: 4121–3.

Harvey, W. T. et al. (2021) 'SARS-CoV-2 Variants, Spike Mutations and Immune Escape', *Nature Reviews Microbiology*, 19: 409–24.

Hatcher, E. L. et al. (2017) 'Virus Variation Resource—Improved Response to Emergent Viral Outbreaks', *Nucleic Acids Research*, 45: D482–90.

Hodcroft, E. B. (2021) *CoVariants: SARS-CoV-2 Mutations and Variants of Interest* <https://covariants.org/> accessed 9 Jun 2023.

Holm, C., Kékicheff, P., and Podgornik, R. eds (2000) *Electrostatic Effects in Soft Matter and Biophysics*. Amsterdam: Kluwer Academic Press.

Huang, Y. et al. (2020) 'Structural and Functional Properties of SARS-CoV-2 Spike Protein: Potential Antivirus Drug Development for COVID-19', *Acta Pharmacologica Sinica*, 41: 1141–9.

International Committee on Taxonomy of Viruses (2021) *Virus Metadata Resource* <https://ictv.global/vmr> accessed 23 Jan 2023.

Javidpour, L. et al. (2021) 'Electrostatic Interactions Between the SARS-CoV-2 Virus and a Charged Electret Fibre', *Soft Matter*, 17: 4296–303.

Jawad, B. et al. (2022) 'Binding Interactions Between Receptor-Binding Domain of Spike Protein and Human Angiotensin Converting Enzyme-2 in Omicron Variant', *The Journal of Physical Chemistry Letters*, 13: 3915–21.

Kim, S. H. et al. (2022) 'Positively Bound: Remapping of Increased Positive Charge Drives SARS-CoV-2 Spike Evolution to Optimize its Binding to Cell Surface Receptors', *chemRxiv*.

Lide, D. R. ed. (2013) *CRC Handbook of Chemistry and Physics*, 94th edn. Boston: CRC press.

Li, S., and Zandi, R. (2022) 'Biophysical Modeling of SARS-CoV-2 Assembly: Genome Condensation and Budding', *Viruses*, 14: 2089.

Loo, L. et al. (2023) 'Fibroblast-Expressed LRRC15 is a Receptor for SARS-CoV-2 Spike and Controls Antiviral and Antifibrotic Transcriptional Programs', *PLOS Biology*, 21: e3001967.

Magazine, N. et al. (2022) 'Mutations and Evolution of the SARS-CoV-2 Spike Protein', *Viruses*, 14: 640.

Moorthy, V. et al. (2020) 'Data Sharing for Novel Coronavirus (COVID-19)', *Bulletin of the World Health Organization*, 98: 150.

Nap, R. J. et al. (2014) 'The Role of Solution Conditions in the Bacteriophage PP7 Capsid Charge Regulation', *Biophysical Journal*, 107: 1970–9.

Nie, C. et al. (2022) 'Charge Matters: Mutations in Omicron Variant Favor Binding to Cells', *ChemBioChem*, 23: e202100681.

Obermeyer, F. et al. (2022) 'Analysis of 6.4 Million SARS-CoV-2 Genomes Identifies Mutations Associated with Fitness', *Science*, 376: 1327–32.

O'Toole, Á. et al. (2021) 'Tracking the International Spread of SARS-CoV-2 Lineages B.1.1.7 and B.1.351/501Y-V2 with Grinch', *Wellcome Open Research*, 6.

——— et al. (2022) 'Pango Lineage Designation and Assignment Using SARS-CoV-2 Spike Gene Nucleotide Sequences', *BMC Genomics*, 23: 1–13.

Pawłowski, P. H. (2021) 'Additional Positive Electric Residues in the Crucial Spike Glycoprotein S Regions of the new SARS-CoV-2 Variants', *Infection and Drug Resistance*, 14: 5099–105.

Pettersen, E. F. et al. (2004) 'UCSF Chimera—a Visualization System for Exploratory Research and Analysis', *Journal of Computational Chemistry*, 25: 1605–12.

Pöyry, S., and Vattulainen, I. (2016) 'Role of Charged Lipids in Membrane Structures—Insight Given by Simulations', *Biochimica Et Biophysica Acta (BBA)—Biomembranes*, 1858: 2322–33.

Pratt, B., Bull, S., and Med, B. M. C. (2021) 'Equitable Data Sharing in Epidemics and Pandemics', *Ethics*, 22: 1–14.

——— et al. (2022b) 'Quantum Chemical Computation of Omicron Mutations Near Cleavage Sites of the Spike Protein', *Microorganisms*, 10: 1999.

Sang, P. et al. (2022) 'Electrostatic Interactions are the Primary Determinant of the Binding Affinity of SARS-CoV-2 Spike RBD to ACE2: A Computational Case Study of Omicron Variants', *International Journal of Molecular Sciences*, 23: 14796.

Šiber, A., Božič, A., and Podgornik, R. (2012) 'Energies and Pressures in Viruses: Contribution of Nonspecific Electrostatic Interactions', *Physical Chemistry Chemical Physics*, 14: 3746–65.

Singh, D., and Yi, S. V. (2021) 'On the Origin and Evolution of SARS-CoV-2', *Experimental & Molecular Medicine*, 53: 537–47.

Telenti, A., Hodcroft, E. B., and Robertson, D. L. (2022) 'The Evolution and Biology of SARS-CoV-2 Variants', *Cold Spring Harbor Perspectives in Medicine*, 12: a041390.

Zandi, R. et al. (2020) 'On Virus Growth and Form', *Physics Reports*, 847: 1–102.

Zhang, W. et al. (2022) 'Structural Basis for Mouse Receptor Recognition by SARS-CoV-2 Omicron Variant', *Proceedings of the National Academy of Sciences*, 119: e2206509119.

Zhou, H.-X., and Pang, X. (2018) 'Electrostatic Interactions in Protein Structure, Folding, Binding, and Condensation', *Chemical Reviews*, 118: 1691–741.